Assessing Item and Scale Differential Functioning using the DFIT Methodology

Leo S. Morales, M.D., Ph.D.

DRAFT NCI Paper

June 10, 2004

Introduction

Methods for assessing differential functioning (DF) refers to statistical methods for examining bias in survey items or scales. An item or scale shows DF if individuals having the same ability or standing on an attribute that belong to different groups have different probabilities of obtaining a similar score on the same item or scale. The effect of items and scales showing DF can be that differences between groups in mean levels or in the pattern of correlations of an item or scale with other variables are artifactual and may be substantively misleading. Various methods grounded in item response theory (IRT) have been described for assessing DF at the item level and scale level.[1,2]

In the present study, we evaluate the Mini Mental Status Examination (MMSE) for differential item functioning (DIF) and differential scale functioning (DTF) using the DFIT framework,[3] that is based on IRT. Whereas most methods for assessing differential functioning focus exclusively on item level evaluations, the DFIT framework provides an approach for simultaneously evaluating DF at both the item level and scale levels. Whether item- or scale-level evaluations are more important to a practitioner will depend on the purpose of the DF analysis. If the analysis is in the context of scale development, in which items are being selected from a pool of potential items, item-level DF or differential item functioning (DIF) may be more important. Typically, items showing DIF are removed. However, if the purpose of analysis is to examine the comparability of individuals belonging to two or more groups as is often the case in disparities research, then scale-level analysis maybe of more interest. Because much of the IRT literature is rooted in educational research, DF at the scale-level is commonly referred to as differential "test" functioning or DTF.

Methods

Subjects. The data for this study came from a case registry study of dementia developed among individuals 65 and over living in 13 census tracks in North Manhattan, New York. Analyses for this study were performed by comparing the item response patterns of those individuals interviewed in Spanish (n=665) and those individuals interviewed in English (n=913).

Measures. The MMSE contains 20 items which in aggregate measure the degree of cognitive impairment in the areas of orientation, attention and calculation, registration, recall and language, as well as the ability to follow verbal and written commands (see Table 1).[4] Recall or short-term memory is measured by asking the respondent to recall a list of three objects memorized earlier. Typical orientation items are "cannot name state s/he is in", "does not know the current year". Attention is measured by asking the respondent to spell the word "WORLD" backwards or to calculate "serial 7's". Language is measured by object naming tasks and by tasks requiring following directions; e.g., folding a piece of paper. In the present study, the "WORLD backwards" and "serial 7's" items were analyzed separately, thus giving a total of 21 items.

Research Design. The analytic procedure used in the present study includes several steps. In the first step, the dimensionality of the Spanish and English MMSE data was assessed by confirmatory factor analyses. We use an IRT model that assumes unidimensional scales. Next, IRT item and person parameters were estimated for the Spanish (focal group) and English (reference group) samples separately using Samejima's two-parameter graded response model[5,6] as implemented in MULTILOG 7.0.[7] Separate estimation of the IRT parameters for the two language groups required that the Spanish group IRT parameters be transformed to the same scale as the English group IRT parameters. To accomplish this, scale linking parameters were estimated. However, because some of the items used in estimating the initial set of linking

parameters were subsequently found to show DIF, an iterative linking procedure was used to estimate linking parameters solely based on items free of DIF.[8-10] Once the linking parameters had been "purged" of items showing DIF, the final set analyses of differential functioning of items and test was performed.

Assessing Differential Functioning. A number of approaches for assessing DF are available (see Teresi, 2004). In this study, we assessed DIF on the basis of two indexes: the non-compensatory differential item functioning (NCDIF) index and compensatory differential item functioning (CDIF) index. Both indexes have been described in detail elsewhere.[3] Basically, the NCDIF index assesses DIF in a particular item under the assumption that the other items in a scale are DIF free. This is the same assumption typically made by other approaches to DIF assessment when anchor items are not specified. For example, when the IRTLRDIF program is used without specifying the a set of anchor items, all items except the item being tested are assumed to be DIF free (meaning that equality constraints are applied across groups for all items except the one being tested).

In contrast to the NCDIF index, the CDIF index allows all items in a scale to show DIF simultaneously. Furthermore, summarizing the item CDIF index values provides an assessment of differential functioning at the scale-level. Because the CDIF index may take on positive or negative values depending on the directionality of the DIF, it is possible for cancellation of DIF to occur at the scale level. For this to occur, one or more items showing DIF in favor of one group are offset by other items showing DIF in favor of the other group. The combination of these two sets of DIF items may have a canceling effect on differential functioning at the scale-level resulting in DIF at the item-level but not the scale-level.

Statistical tests for the NCDIF index and CDIF index have not been developed. Instead, cutoff values similar to those used in confirmatory factor analysis (e.g., Comparative Fit Index) have been established for each index on the basis of simulation studies. The cutoff for the NCDIF index varies by the number of response categories for an item. For a dichotomous item, the cutoff value is 0.006; for items with 4 response categories, the value is 0.054; and for items with 6 categories, the value is 0.150. The CDIF index cutoff value for the scale was 0.558.

Results

Descriptive Statistics. Table 1 shows descriptive statistics for the Spanish and English data for the MMSE including items means and standard deviations, item-test correlation coefficients corrected for overlap, and coefficient alpha for the MMSE scale if the item were dropped from the scale.

Scale Dimensionality. The confirmatory factor analyses showed that a one factor solution adequately represented the data in the Spanish and English samples. For the Spanish sample, a one factor solution resulted in a Tucker-Lewis Index (TLI) of 0.99, a Comparative Fit Index (CFI) of 0.97, a Root Mean Square Error of Approximation (RMSEA) of 0.06, and a Standardized Root Mean Square Residual (SRMR) of 0.06. For the English sample, a one factor solution resulted in a TLI of 0.99, a CFI of 0.97, a RMSEA of 0.06, and a SRMR of 0.05.

Item-Level Differential Functioning. Nine items were found to show DIF on the basis of NCDIF index values that exceeded the recommended cutoff value (See Table 3). These items included the items that ask the respondent to name the correct season (NCDIF=0.060), name the correct day of the month (NCDIF=0.030), name the correct city (NCDIF=0.112) and state(NCDIF=0.025), name two nearby streets (NCDIF=0.005)[1], recall three

[1] The NCDIF value for this item exceeded the recommended cutoff in an earlier iteration and therefore was deemed to show DIF in spite of an acceptable NCDIF value in the final iteration.

objects(NCDIF=0.106), repeat the phrase *no ifs, no ands, no buts* (NCDIF=0.098), follow the command, "close your eyes" (NCDIF=0.010), and follow the command, "take the paper in your right hand, fold the paper in half with both hands, and put the paper down in your lap" (NCDIF=0.087).

Scale-Level Differential Functioning. The English and Spanish versions of the MMSE were not found to show differential functioning (scale CDIF=0.215) despite nine of the individual items showing DIF (See Table 3). As discussed previously, this is possible because of the compensatory nature of DIF at the item-level.

Discussion

Our results show that 9 of the 21 items that make up the MMSE show significant DF. The items showing DIF include items that ask the respondent to name the correct season, name the correct day of the month, name the correct city and state, name two nearby streets, recall 3 objects, repeat the phrase *no ifs, no ands, no buts*, follow the command, "close your eyes", and follow the command, "take the paper in your right hand, fold the paper in half with both hands, and put the paper down in your lap". As a result, the English and Spanish versions of these items may not be comparable and may result in incorrect conclusions when used on an item-by-item basis.

Offsetting effects among the items in the MMSE, however, resulted in a non-significant degree of differential functioning at the scale level. Inspection of the CDIF indices for the English and Spanish MMSE items showed that some items had a positive CDIF index value while others had a negative CDIF value. These opposing forms of DIF resulted in a non-significant degree of DF at the scale level. As a result, assessments of cognitive functioning

among respondents to the English and Spanish versions of the MMSE are comparable at the scale level.

The psychometric analyses conducted for this study do not provide many clues as to why certain items function differently across the groups analyzed. The reasons for differential functioning can be multiple and require further research to ascertain. In general, differential functioning can occur for a number of reasons including differences in the meaning of items, differences in the cognitive processes used in responding, differences in the appropriateness of response scales, problems in translation or interpretation and the appropriateness of data collection procedures.[11] Qualitative methods such as focus groups and cognitive interviews are necessary to assess the reasons underlying item differential functioning.

REFERENCES

1. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice 1998; 17:31-44.
2. Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement 1993; 17:297-334.
3. Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement1995; 19, 353-368.
4. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State: A practical guide for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975;12:189-98.
5. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph 1969; 17.
6. Samejima F. A new family of models for the multiple-choice item. Research Report No. 79-4, Department of Psychology, University of Tennessee; 1979.
7. du Toit M. IRT from SSI. Lincolnwood, IL: Scientific Software International; 2003.
8. Ellis BB, Mead AD. Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. Educational and Psychological Measurement 2000; 60(5):787-807.
9. Candell GL, Drasgow F. An iterative procedure for linking metrics and assessing item bias in item response theory. Applied Psychological Measurement 1988; 12(3):253-260.
10. Collins WC, Edwards JE, Raju NS. Assessing differential functioning in a satisfaction scale. Journal of Applied Psychology 2000; 85(3):451-61.

11. Stewart AL, Napoles-Springer AM. Advancing health disparities research: can we afford to ignore measurement issues? Medical Care 2003; 41:1207-20.

12. TABLE 1. Mini mental status exam (MMSE) Item content.

Item	Abbreviated Item Content
MMSE 1	Doesn’t state year correct
MMSE 2	Doesn’t state season correct
MMSE 3	Doesn’t state correct day of month
MMSE 4	Doesn’t state correct day of week
MMSE 5	Doesn’t state correct month
MMSE 6	Doesn’t state correct state
MMSE 7	Doesn’t state correct city
MMSE 8	Doesn’t state two nearby streets
MMSE 9	Doesn’t state correct floor
MMSE 10	Doesn’t correctly identify type of place
MMSE 11	Apple, table, penny errors
MMSE 12	Errors in serial 7 subtraction
MMSE 13	Errors in spelling WOLRD backwards
MMSE 14	Errors in recalling 3 objects
MMSE 15	Doesn’t name pencil
MMSE 16	Doesn’t name wristwatch
MMSE 17	Errors in repeating phrase
MMSE 18	Doesn’t close eyes
MMSE 19	Errors following instructions with paper
MMSE 20	Error writing complete sentence
MMSE 21	Error copying design

TABLE 2. Descriptive Statistics for English and Spanish MMSE Items

		English (Alpha=0.89)				Spanish (Alpha=0.87)			
Item	**Categories**	**Mean**	**SD**	**Item-Test**	**Alpha***	**Mean**	**SD**	**Item-Test**	**Alpha***
MMSE 1	2	0.30	0.46	0.69	0.88	0.26	0.44	0.68	0.86
MMSE 2	2	0.28	0.45	0.67	0.89	0.42	0.49	0.52	0.86
MMSE 3	2	0.50	0.50	0.54	0.89	0.37	0.48	0.58	0.86
MMSE 4	2	0.28	0.45	0.67	0.89	0.24	0.42	0.62	0.86
MMSE 5	2	0.28	0.45	0.66	0.89	0.25	0.43	0.64	0.86
MMSE 6	2	0.20	0.40	0.61	0.89	0.43	0.50	0.43	0.86
MMSE 7	2	0.12	0.33	0.66	0.89	0.16	0.37	0.63	0.86
MMSE 8	2	0.28	0.45	0.68	0.89	0.27	0.44	0.65	0.86
MMSE 9	2	0.19	0.39	0.68	0.89	0.13	0.34	0.62	0.86
MMSE 10	2	0.21	0.41	0.67	0.89	0.14	0.34	0.62	0.86
MMSE 11	4	0.35	0.91	0.62	0.88	0.22	0.75	0.62	0.86
MMSE 12	6	2.98	1.86	0.56	0.90	2.88	1.79	0.50	0.88
MMSE 13	6	2.21	1.91	0.71	0.89	2.25	1.88	0.64	0.87
MMSE 14	4	2.14	1.11	0.48	0.89	1.86	1.14	0.45	0.86
MMSE 15	2	0.11	0.31	0.64	0.89	0.05	0.23	0.58	0.86
MMSE 16	2	0.12	0.33	0.68	0.89	0.06	0.24	0.60	0.86
MMSE 17	2	0.39	0.49	0.48	0.89	0.11	0.31	0.54	0.86
MMSE 18	2	0.18	0.39	0.69	0.89	0.19	0.39	0.66	0.86
MMSE 19	4	0.83	1.00	0.72	0.88	0.54	0.87	0.63	0.85
MMSE 20	2	0.29	0.46	0.69	0.89	0.30	0.46	0.61	0.86
MMSE 21	2	0.56	0.50	0.49	0.89	0.54	0.50	0.44	0.86

*Alpha if item deleted form scale.

TABLE 3. Final DFIT Results for MMSE English versus Spanish.

Item	NCDIF Cutoff	NCDIF	CDIF
MMSE 1	0.006	0.001	0.005
MMSE 2	0.006	0.060+	-0.074
MMSE 3	0.006	0.030+	0.074
MMSE 4	0.006	0.000	0.000
MMSE 5	0.006	0.003	0.004
MMSE 6	0.006	0.112+	-0.091
MMSE 7	0.006	0.025+	-0.011
MMSE 8	0.006	0.005++	-0.001
MMSE 9	0.006	0.001	0.002
MMSE 10	0.006	0.003	0.002
MMSE 11	0.054	0.008	0.000
MMSE 12	0.150	0.011	0.032
MMSE 13	0.150	0.021	-0.024
MMSE 14	0.054	0.106+	0.127
MMSE 15	0.006	0.002	0.004
MMSE 16	0.006	0.004	0.006
MMSE 17	0.006	0.098+	0.076
MMSE 18	0.006	0.010+	-0.003
MMSE 19	0.054	0.087+	0.095
MMSE 20	0.006	0.003	-0.014
MMSE 21	0.006	0.000	0.006
DTF Cutoff	0.558		
Scale DTF	0.215		

+ NCDIF value exceeded NCDIF cutoff in final run of the DFIT program.
++NCDIF value exceeded NCDIF cutoff in earlier run of the DFIT program.